Prominence and coherence in a Bayesian theory of pronoun interpretation

Prominence and coherence in a Bayesian theory of pronoun interpretation

Prominence and Coherence in a Bayesian Theory of Pronoun Interpretation

Andrew Kehler

University of California, San Diego

Hannah Rohde

University of Edinburgh

Abstract

A standard assumption in linguistic and psycholinguistic research on pronoun use is that production and interpretation are guided by the same set of contextual factors. Kehler et al. (2008) and Kehler & Rohde (2013) have argued instead for a model based on Bayesian principles in which pronoun production is insensitive to a class of semantically- and pragmatically-driven contextual biases that have been shown to influence pronoun interpretation. Here we evaluate the model using a passage completion study that employs a subtle contextual manipulation to which traditional analyses are insensitive, specifically by varying whether or not a relative clause that modifies the direct object in the context sentence invites the inference of a cause of the event that the sentence denotes. The results support the claim that pronoun interpretation biases, but not production biases, are sensitive to this pragmatic enrichment, revealing precisely the asymmetry predicted by our Bayesian Model. A correlation analysis further establishes that the model provides better estimates of measured pronoun interpretation biases than two competing models from the literature.

Keywords: Pronoun interpretation, discourse coherence, pragmatic enrichment, Bayesian models

April 2, 2018

1. Introduction

A common wisdom, one assumed in the literature on pronominal reference for decades, is that there is a unified notion of entity PROMINENCE that underlies pronoun usage. This notion of prominence (alternately referred to as SALIENCE, being in PSYCHOLOGICAL FOCUS, being THE CENTER OF ATTENTION, and so forth) will determine when a speaker will choose to use a pronoun on the one hand, and hence be used by the addressee to successfully interpret the reference on the other. On this assumption, the task for discourse researchers is then clear: One merely needs to identify the different factors that contribute to entity prominence. Many such factors have been posited, including grammatical role [11], grammatical parallelism [42], thematic role [43, 3], information structure [18, 37], semantics [29], and world knowledge [22], among others.

A central goal of this paper is to disabuse the reader of this assumption. We do this by evaluating a proposal put forth by Kehler et al. [25] and Kehler and Rohde [26] that states that the relationship between pronoun interpretation and production is suitably cast in Bayesian terms, and further that the types of factors that condition the likelihood term of Bayes' Rule (the production bias) and the prior (the bias toward entity next mention) are different. This entails the counterintuitive conclusion that a set of factors that the addressee will use in resolving the referent will not be taken into account by the speaker when deciding whether to produce a pronoun. We evaluate the causal structure that underlies the proposal with a novel passage completion experiment which confirms the predictions of the analysis. We further show that the biases revealed by the experimental results are more highly correlated with the predictions of the Bayesian Model than those of two other prominent models. Consequences for the notion of prominence as it relates to the interpretation of referring expressions are discussed.

2. Background

2.1. Previous Work on Pronoun Interpretation

The idea that a unified notion of prominence underlies reference interpretation and production has a substantial history in theoretical and experimental linguistics over the last several decades. The underlying thought is intuitive: At any particular time during a discourse, certain entities are being attended to more than others by the conversational participants, and are likewise more likely to be re-mentioned as the discourse ensues. As such, speakers can get away with using reduced, if albeit semantically ambiguous, pronominal forms to refer to such entities, under the assumption that they and their hearers are relying on the same cues to entity prominence. Following Rohde and Kehler [37], we will refer to accounts that assume such a relationship as MIRROR MODELS, so named to capture the idea that pronoun interpretation biases mirror the biases that underlie the speaker's decision to produce a pronoun.

The relationship between prominence and referential form has been posited in a number of theories that seek to explain the contextual conditions under which different referring expressions will be used.[1] For instance, Givón [17] associated a speaker's choice of referring expression with a gradient scale of topicality that entities can hold. According to his analysis, reduced referential expressions (zeros, unstressed pronouns) correspond to those entities with the highest topicality. In a similar spirit, Ariel [1] proposed an ACCESSIBILITY HIERARCHY with which the production and interpretation of referring expressions are said to correlate. In her account, the higher the degree of accessibility that the intended referent holds, the lower the amount of lexical material that will be used in the referring expression chosen by the speaker. As a particularly reduced form of reference, pronouns are thus associated with a high presumed degree of accessibility. Finally, Gundel et al. [21] propose a GIVENNESS HI-

[1] For an accessible overview of the literature on reference and various notions of prominence, see Arnold [4].

ERARCHY containing six possible cognitive statuses that a referent may hold with respect to the hearer's knowledge state and mental model of the discourse. These statuses serve as the conventional meanings for different types of referring expressions across languages; pronouns are associated with the strongest status IN FOCUS, which they characterize as holding when "the referent is...at the current center of attention" (p. 279). Whereas there are differences with respect to the pragmatic notions underlying the scales with which these authors associate referential form, all posit that speakers and hearers take into account the correspondence between form of reference and relative status of the referent on these pragmatic scales in their productions and interpretations respectively. That is, in selecting a particular referring expression, the speaker signals that she assumes that the associated felicity conditions are met, and hence the hearer is expected to utilize those same felicity conditions when identifying a speaker's intended referent.

Whereas the analyses just surveyed have focused on the respective topicality, accessibility, and cognitive status of referents, others have focused on referent predictability. For instance, Arnold's Expectancy Hypothesis [2, 3] explicitly links pronoun interpretation to the hearer's expectations about what entities are likely to be mentioned at the point in the discourse at which a pronoun is encountered. Many factors potentially influence predictability on her model; some that are familiar from the literature on pronoun interpretation were listed in Section 1. The important difference between her model and many previous ones is the relevance of predictive, top-down processing: Since these factors are driving expectancies regarding to-be-mentioned entities, the majority of the hearer's work occurs during the normal course of (predictive) discourse comprehension. At the time a pronoun is subsequently encountered, the hearer can simply match the pronoun against his expectations (taking into consideration morphosyntactic constraints such as gender, number, person, and so forth). Although Arnold's model can also be seen as a type of Mirror Model, it will be useful to distinguish accounts like hers and hence refer to them as EXPECTANCY MODELS, in recognition of the explicit link between prominence and predictabil-

ity captured by the account.

The explicit tie between pronominalization and predictability is relevant to larger controversies concerning communicative efficiency as a design principle of language. For instance, according to the Uniform Information Density (UID) hypothesis [31], language will be most efficient when speakers seek to maximize communicative channel capacity. This requires that a relatively constant rate of information transmission be maintained, which can only be achieved if articulatory effort is kept proportional to the amount of information conveyed, a quantity that is inversely proportional to the predictability of the message. The hypothesis thus predicts that reduced referring expressions like pronouns should be used to refer to entities that are highly predictable in context. Tily and Piantadosi [45] investigated this question by presenting naturally-occurring texts piece-by-piece to participants, who then tried to guess what entity would be mentioned next from those that had been mentioned before (along with a "something new" category for previously unmentioned entities). Their results revealed a strong correlation between predictability and form of reference, whereby participants were more likely to correctly predict the next-mentioned entity in cases in which the text's author had ultimately chosen to use a pronoun as compared to when longer definite noun phrases were used.

2.2. A Complication for the Models

The idea that there is a coherent notion of entity prominence that mediates between production and interpretation is therefore highly compelling. However, there are also results from the literature that call the idea into question. We illustrate these issues with examples involving so-called IMPLICIT CAUSALITY (IC) verbs, as much of the literature to be discussed from this point on has utilized them, as does the new experiment we present. IC verbs are undoubtedly the most well-studied verb class in the psycholinguistics of pronoun interpretation literature since the seminal papers of Caramazza and colleagues in the 1970s [15, 9, 8, 5, 32, 29, 25, inter alia]. Such verbs are so-called because they are said to impute causality to one of the participants associated with the event

they denote, which in turn affects subsequent referential biases. For instance, if participants in a passage completion task are presented with a prompt like (1a),

(1) a. Amanda amazed Brittany because she _____

b. Amanda detested Brittany because she _____

the large majority of completions will point to Amanda as the pronominal referent. After all, Amanda must be amazing, and hence one expects to hear why. Because causality is imputed to the subject, verbs like *amaze* are called SUBJECT-BIASED IC VERBS. If participants are given a prompt like (1b), on the other hand, the large majority of completions will point to Brittany as the pronominal referent. After all, Brittany must be detestable, and hence one expects to hear why. Because causality is imputed to the object, verbs like *detest* are called OBJECT-BIASED IC VERBS. The existence of IC biases has been replicated repeatedly, and is hence one of the bedrock results in the field.

Stevenson et al. [43] reported on a series of passage completion experiments that investigated pronoun biases across eight distinct context types, including two subclasses of IC verb (specifically, Stimulus-Experiencer and Experiencer-Stimulus verbs). In the manipulation for the experiment of interest here (their Experiment 1), context sentences ended with a full stop, and the follow-ons varied between including or not including a pronoun in the prompt:

(2) a. Amanda detested Brittany. She _____

b. Amanda detested Brittany. _____

Unlike (2a), the free prompt condition (2b) allows participants to pick their own referring expressions for the first-mentioned entity. Stevenson et al. found two results of interest. First, across all eight context types, there were a greater number of references to the previous subject in the pronoun prompt condition than in the free prompt condition. Crucially, this did not always result in an overall pronoun bias toward the subject in the pronoun prompt condition: For instance, for object-biased IC verbs as in (2a), the overall pronoun bias was still toward the object. Instead, the key finding is that the occurrence of a pronoun

in the prompt always shifted the distribution of references toward the subject compared to when no pronoun was provided – that is, for prompts like (2b), the first-mention bias toward the object was even stronger.

Stevenson et al.'s second finding was that, in their free prompt conditions across all stimulus types, participants' choice of referential form for the first-mentioned entity was heavily biased towards a pronoun when the referent was the previous subject, and likewise towards a name when the referent was a non-subject. This result may at first seem contradictory: If participants have a clear preference to use pronouns to refer to the previous subject and names to refer to non-subjects, why would the pronoun interpretation bias ever be toward a non-subject, as was the case for prompts like (2a)? Hence we have initial evidence for a dissociation between production biases and interpretation biases, casting doubt on whether the same notion of prominence could underlie both processes.

Other studies similarly cast doubt on the idea of a unified notion of entity prominence. In an analysis of the behavior of null and overt pronouns in Greek, Miltsakaki [33] found that whereas participants in a passage completion study tended to refer to the semantically-focused referent in a free-prompt condition, they were reluctant to use a pronoun to refer to such referents when they were in object position, opting for an overt pronoun instead. The production of null pronouns, instead, was heavily biased toward the referent occupying subject position, regardless of the semantic bias. This result led Miltsakaki to conclude that prominence and choice of referring expression can be sensitive to different properties of discourse entities. Other work indicates that the ways in which different forms of reference are interpreted within a language are sensitive to different factors that confer entity prominence. For instance, Kaiser and Trueswell [24] reported on a study that compared the behavior of the Finnish pronoun *hän* (s/he) with the demonstrative *tämä* (this) with respect to the grammatical role and word order position of possible antecedents. The results revealed that *hän* is sensitive primarily to syntactic role (preferring the subject of the previous sentence) regardless of word order, whereas *tämä* is sensitive mainly to word order (preferring postverbal referents, with only a smaller secondary effect

favoring objects over subjects). Their results lead them to conclude that there is no single notion of entity prominence that governs all referential form usage; instead different referential expressions exhibit varying degrees of sensitivity to different contextual factors.

2.3. The Bayesian Model

At this point we have evidence for an asymmetry between reference interpretation and production [43, 33], but no model that makes quantitative predictions about the relationship between the two. Kehler et al. [25] offered an analysis that addresses this need. Specifically, they proposed that the relationship between production and interpretation is Bayesian, as shown in equation (3).

$$(3)\quad P(\textit{referent} \mid \textit{pronoun}) = \frac{P(\textit{pronoun} \mid \textit{referent})\ P(\textit{referent})}{\sum_{\textit{referent} \in \textit{referents}} P(\textit{pronoun} \mid \textit{referent})\ P(\textit{referent})}$$

The term $P(\textit{referent} \mid \textit{pronoun})$ represents the interpretation bias: The probability, given that a pronoun has occurred, of it being used by the speaker to refer to a particular referent. On the other hand, the term $P(\textit{pronoun} \mid \textit{referent})$ represents the production bias: The probability, assuming that a particular entity is being referred to, that the speaker would have used a pronoun to refer to it. Bayes' Rule says that these biases are not mirror images of each other, but instead are related by the prior $P(\textit{referent})$, which represents the NEXT-MENTION bias: The probability that a particular referent will get mentioned next regardless of the referring expression used.[2] On this model, therefore, comprehenders reverse-engineer the referential intentions of the speaker by combining their estimates of the speaker's production biases with their prior expectations about the probability that particular entities will be mentioned next. Equation (3) thus explains why there is nothing contradictory about having both a strong

[2]The denominator of equation (3) is simply the probability that a pronoun is the form of reference chosen by the speaker ($P(\textit{pronoun})$), which can be computed by calculating the value in the numerator for each referent that is compatible with the pronoun and summing the result. This has the effect of normalizing the probabilities so that they sum to 1.

production bias toward pronominalizing the previous subject (and not pronominalizing non-subjects) and yet a lack of a subject bias in interpretation, as long as the prior $P(\textit{referent})$ points strongly enough away from the subject referent, as it does for object-biased IC verbs, for example.[3]

Kehler et al.'s Bayesian Model comes in two varieties. As it stands, equation (3) says only that the relationship between pronoun interpretation and pronoun production follows Bayesian principles, without further specifying the types of contextual factors that affect each term in the numerator. We refer to this claim as the WEAK form of the Bayesian Model. The weak analysis therefore simply predicts that if independent estimates of the prior, likelihood, and posterior probabilities were obtained, equation (3) would approximately hold. Kehler et al. also suggested, however, that the two terms in the numerator of equation (3) are conditioned by different types of contextual factors. We call this the STRONG form of the Bayesian Model. On the one hand, they noted that the data they surveyed suggested that the factors that condition the next-mention bias $P(\textit{referent})$ are primarily semantic (e.g., verb type) and pragmatic (e.g., coherence relations; see next section). On the other hand, the factors that condition the production bias $P(\textit{pronoun} \mid \textit{referent})$ appear to be grammatical and/or information structural (e.g., based on grammatical role obliqueness or topichood, both of which amount to a preference for sentential subjects). Considering this in light of the asymmetry between production and interpretation captured by equation (3), the strong hypothesis makes a striking prediction: That the speaker's decision about whether or not to pronominalize a reference will be insensitive to the semantically- and pragmatically-driven contextual factors that in part determine the comprehender's interpretation bi-

[3]The Bayesian Model, of course, is intended to apply to all contexts in which pronominal reference could occur. The best scenarios for testing the analysis and comparing it to others, however, are provided by contexts in which the production bias and prior are expected to favor different discourse entities, since it is in such contexts that the account predicts the greatest dissociation between production and interpretation. This is one reason we utilize object-biased IC contexts in the current study.

ases. This hypothesis is surprising because it violates the intuition, represented in the prior work we surveyed in Section 2.1, that speakers will pronominalize mentions of referents in just those cases in which their comprehenders would be expected to interpret the pronouns to those same referents.

As unintuitive as this may seem, the results of several recent passage completion studies have provided preliminary support for this prediction (Rohde, 2008, Experiments V–VII; Fukumura and van Gompel, 2010). We briefly describe one reported on by Rohde and Kehler (2014; see also Rohde, 2008, Experiment VI) which, unlike the other studies, examined pronoun usage in referentially-ambiguous contexts. Rohde and Kehler's experiment consisted of a 3x2 design that crossed context type with prompt type (4).

(4) a. John infuriated Bill. (He) _____

b. John scolded Bill. (He) _____

c. John chatted with Bill. (He) _____

The 3-way context manipulation compared subject-biased IC contexts (4a), object-biased IC contexts (4b), and non-IC contexts (4c), and the 2-way prompt condition compared free prompts with pronoun prompts. The production bias was measured by examining the choice of referring expression (pronoun or name) that participants employed for the first-mentioned referent in the free prompt condition, and the interpretation bias was measured using their completions in the pronoun prompt condition. As expected, the interpretation bias varied with the IC bias across the context types, with subject mentions being most frequent for (4a), least in (4b), and (4c) being in between. However, this context difference did not affect rate of pronominalization in the free prompt condition. Instead, only grammatical role mattered, with participants pronominalizing subject references far more often than non-subject ones. To put a fine point on this: Participants were no more likely to pronominalize an object mention in an object-biased IC context like (4b) than in a subject-biased IC one

like (4a), and similarly no more likely to pronominalize a subject mention in a subject-biased context than in an object-biased one.

2.4. Causal Coherence

The strong Bayesian Model makes a wide variety of predictions not only about what factors will influence pronoun interpretation, but also about the causal structure that underlies their influence. The two components that feed interpretation biases on the model are captured in the simple graphical representation shown in Figure 1. The specific prediction tested by the experiment

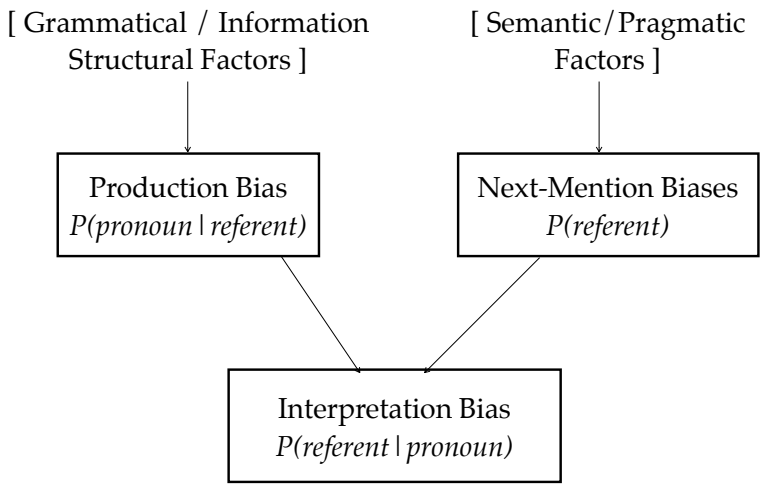

Figure 1: A Basic Graphical Model

described at the end of the previous section varied a semantic factor – IC verb bias – to test the hypothesis that it would have an effect on interpretation biases imposed indirectly through an effect on the prior, while having no effect on the production bias. The hypothesis was confirmed, providing preliminary support for the account.

Ultimately, the account predicts that *anything* that affects the prior, no matter how seemingly independent of pronominalization, should nonetheless show an influence on pronoun interpretation. At the same time, relatively few,

if any, of such factors should also affect pronoun production. Our goal in this paper is to find a pragmatic factor that, by way of a chain of linguistically-motivated causal dependencies, would be predicted by the model to have this differential effect on interpretation and production. The manipulation we use brings together the referential biases associated with IC verbs that we have discussed with two other semantic and pragmatic facts, described here.

The first fact is that IC verbs are associated not only with strong referential biases, but also another type of bias that bears on the COHERENCE RELATIONS that hold between clauses. Kehler et al. [25] conducted a free-prompt passage completion study that used a variant of the three-way context manipulation shown in (4a-c). The results revealed that participants completed passages such as (4a-b) with sentences that participated in an EXPLANATION coherence relation – i.e., in which an utterance describing an event or state is followed by one describing a cause of or reason for it – approximately 60% of the time. This figure compared with only 24% for a control group of non-IC contexts (4c). This result accords with intuitions: Upon hearing *John amazed Mary*, it seems likely that the addressee will wonder *Why?*, and thus expect to hear an answer. On the other hand, upon hearing *John saw Mary*, it seems less likely that the addressee will wonder *Why?*, and instead expect an answer to a different question, for example, *What happened next?*.

The second fact utilized by the current study is that causal inferences can be established between contents arising not only from adjacent sentences, but also from different constituents *within* a sentence. Consider (5a-b):

(5) a. The boss fired the employee who was embezzling money.

b. The boss fired the employee who was hired in 2002.

In a typical context, a comprehender is likely to infer from sentence (5a) not only that the employee was fired and was embezzling money, but that he was fired because he was embezzling money. Note this inference, which relates the meanings of the matrix verb and the relative clause (RC) that modifies the

direct object, is merely invited (i.e., not entailed) and is hence defeasible: (5a) could be felicitously followed with a sentence such as *The reason the employee was fired is that he never showed up to work on time.* In contrast, sentence (5b) will not generally be enriched to a causal interpretation, that is, it will not be taken to imply that being hired in 2002 was the cause for the firing. Here the RC is understood to simply be identificational, i.e. to restrict the reference of the noun phrase to which it attaches.

Rohde et al. [38] previously examined the role of such causal inferences in RC processing. Their studies utilized the three facts we have just surveyed – i.e. the strong referential biases associated with IC verbs, the fact that IC verbs create an expectation for an ensuing explanation, and the fact that RCs can invite the inference of an explanation – in order to test whether discourse biases can influence syntactic attachment decisions. To see the logic, consider the relative clauses in (6–7) with the sample follow-ons in (a-b):

(6) John babysits the children of the musician who...

a. ...is a resident of La Jolla. [low]

b. ...are students at a private school. [high]

(7) John detests the children of the musician who...

a. ...lives in La Jolla. [low]

b. ...are arrogant and rude. [high]

Note that there is a temporary attachment ambiguity at the time that the head of the RC (i.e., *who*) is encountered: The RC could attach to the HIGH NP, headed by *children*, or the LOW NP, headed by *musician*. It has been widely documented that English has a low-attachment bias for RCs [13, 10, 12, inter alia]. The preference for low attachment therefore predicts uniform biases across (6-7); for instance, in a passage completion experiment, one would expect to see more low-attaching completions (6a-7a) than high-attaching ones (6b-7b).

One would likewise expect the RC verb *lives* in (6a-7a), which agrees in number with the lower NP, to be easier to process on-line than the verb *are* in (6b-7b), which agrees with the higher NP.

Rohde et al.'s analysis, however, predicted a difference between (6–7), based on the fact that the matrix clauses differ in the verb employed: The verb *detests* is an object-biased IC verb, whereas *babysits* is non-IC. Rohde et al. asked what one would expect to happen if comprehenders are able to utilize the three aforementioned facts when making a syntactic attachment decision. Specifically, if object-biased IC verbs like *detest* (i) generate a greater-than-usual pragmatic expectation for an ensuing explanation, (ii) comprehenders are implicitly aware that RCs can serve the pragmatic function of describing such an explanation, and (iii) comprehenders have a pragmatic expectation that an ensuing explanation is most likely to be about the direct object, then we might expect a greater bias for the RC to attach to the direct object – which, crucially, is the HIGH attachment point for the RC – in (7) as compared to non-IC verbs as in (6). This prediction results from the fact that non-IC verbs do not create the same expectation that an explanation will ensue, nor do they have as strong of a referential bias towards their direct objects.

This reasoning only goes through, of course, if the three types of pragmatic information described in (i)-(iii) above are utilized in concert during the normal course of syntactic processing. But this is exactly what Rohde et al. found. In a passage completion experiment, participants wrote more continuations that attached high in the IC condition than the non-IC condition. Further, in a reading time experiment, high-attaching RCs were read more quickly that low-attaching RCs in IC contexts, whereas the reverse pattern held for non-IC contexts. In fact, high-attaching RCs in the IC condition were read faster than any of the other three conditions.

3. Study

The experimental results just discussed demonstrate how a seemingly subtle aspect of passage understanding – a causal enrichment – can affect interpretation in empirically measurable respects, in this case by way of biasing where RCs are likely to attach during syntactic processing. In the study that follows, we use a similar manipulation to evaluate the effects of inferred causes in the domain of discourse interpretation, specifically on pronoun interpretation. Example stimuli are shown in (8):

(8) a. The doctor reproached the patient who never takes her medicine.
She _____
[ExplanationRC, PronounPrompt]

b. The doctor reproached the patient who came in at 3pm.
She _____
[NoExplanationRC, PronounPrompt]

c. The doctor reproached the patient who never takes her medicine.
_____
[ExplanationRC, FreePrompt]

d. The doctor reproached the patient who came in at 3pm.
_____
[NoExplanationRC, FreePrompt]

Importantly, accounts of pronoun interpretation that appeal primarily to surface-level characteristics of the context (first-mention, subject assignment, grammatical role parallelism, and so forth) find little to distinguish (8a-b): The verb and event participants are mentioned from the same grammatical positions; indeed the only difference is the RC that attaches to the direct object.

The Bayesian Model does predict a difference, however, based on an interconnected sequence of referential and coherence-driven interdependencies, as illustrated in the elaborated graphical model shown in Figure 2. First, we

manipulate RC type because we predict that participants will write fewer Explanation continuations in (8a) and (8c) than (8b) and (8d) respectively. This is the case because, although previous work suggests that we should expect an explanation to follow a sentence with an IC verb, the RCs in (8a) and (8c) already provide one. Hence participants should be more likely to continue with a different coherence relation [41, 25, 7].

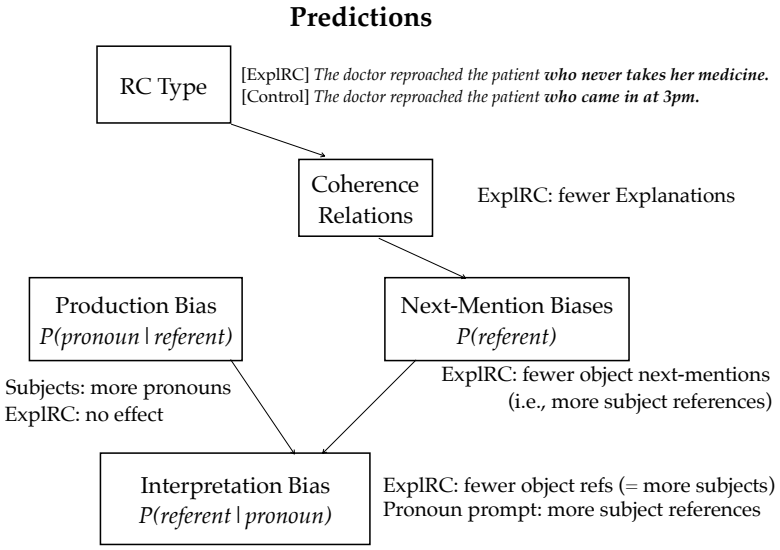

Figure 2: Predictions as a Graphical Model

Second, a difference in the percentage of Explanation continuations is in turn predicted to yield a difference in next-mention biases ($P(referent)$). This is due to the fact that the majority of coherence relations that a participant could employ besides Explanation are more strongly biased to the subject in object-biased IC contexts [25]. That is, whereas even the object bias for Explanation is merely a tendency – (8b) could still be followed with an Explanation continuation with the pronoun referring to the subject, as in *He didn't like the employee's attitude* – the bias toward the object is stronger than that for most other relations in these contexts. Thus, a reduction in the number of Explana-

tion continuations for (8d) compared to (8c) should lead to a greater number of subject biased relations, and hence fewer next mentions of the object.

Third, the analysis predicts that rates of pronoun production ($P(\textit{referent}|\textit{pronoun})$) as measured in the FreePrompt condition (8c-d) should only be affected by grammatical role (favoring pronominalizations of the subject), and hence display no interaction with RC type (with participants no less likely to pronominalize subject mentions, and likewise no more likely to pronominalize object mentions, in 8d than 8c). This accords with the prediction of the strong Bayesian Model that semantic and pragmatic factors assert influence on the next-mention bias but not the production bias.

Finally, RC type is predicted to affect interpretation biases as measured in the PronounPrompt condition, with a greater number of object interpretations in (8b) than (8a), since $P(\textit{referent}|\textit{pronoun})$ is determined in part by next-mention expectations ($P(\textit{referent})$). As such, we predict the RC manipulation to have an effect on pronoun interpretation, but not production. Since interpretation biases are also determined in part by production biases, an effect of grammatical role favoring subjects is also predicted in (8a-b) compared to their FreePrompt counterparts in (8c-d).

3.1. Participants

In service of obtaining a heterogeneous group of self-reported native speakers of English, participants were recruited using Mechanical Turk [34, 16]. Of the 40 total participants, 20 were female, 19 male, and one unspecified. Ages ranged from 21 to 60 years old (mean 32.5), and levels of education ranged from high school through having a Master's degree (30 had at least some college training, and 16 had a 4-year degree or higher).

All participants were located in the United States, with a broad range of locales represented. Because we wanted to use monolingual speakers but didn't want people to be incentivized to lie about their native speaker status, we allowed anyone with a US IP address to participate (and paid everyone). We then only used the data from participants who answered "no" to a question in

a background questionnaire that asked whether any language besides English was spoken at home before the age of 6.

3.2. Design

A 2x2 design was employed using stimuli like (8a-d), which varied RC type (ExplanationRC or NoExplanationRC) and prompt type (PronounPrompt or FreePrompt). The experiment employed 24 target items and 36 fillers. The fillers were included to conceal the target manipulation; context sentences varied between sentences with only one event participant (*Alex rode his bike to work this morning*), two participants of different gender (*Drew chatted with Susan for an hour after the parade*), and two participants of the same gender but without an IC verb (*Stacy played video games with Tanya all afternoon*). Free prompts, pronoun prompts, and connective prompts (*As a result,* _____) were used. Three 'catch' fillers that employed prompts that suggested an obvious continuation were also included to ensure that participants were paying attention (*Caleb did all the cooking for the BBQ even though he hates BBQ. He prefers mac 'n* _____). All participants completed these trials in a manner consistent with paying attention to the task and hence no participant's data was eliminated from analysis as a result.

Context sentences for the target items always used object-biased IC verbs in the matrix clause. Discourse continuations were collected via a web-based interface that participants accessed from their own computer. Each item was presented on a page by itself with a text box in which participants were instructed to write their continuation. Two pieces of clip art were also displayed that indicated the gender of each event participant (always the same for both event participants for target items, so that reference in the PronounPrompt condition would be ambiguous). The task took less than an hour. Participants saw each stimulus in a single condition in a fully balanced design.

Two judges who were blind to the hypotheses annotated the data for coherence relations (Explanation or Other), first-mentioned referent (Subject, Object, or Other), and form of reference in the FreePrompt condition (Pronoun or

Other). For determining whether the coherence relation was Explanation, annotators where asked to consider whether the eventuality described in the second sentence provides an explanation for the eventuality described in the first, and were offered two tests to apply: (i) whether the second sentence answers the question *Why?* with respect to the first, and (ii) whether the connective *because* could be felicitously used to connect the sentences without changing the construal of the passage on its most natural interpretation. So as to not bias the annotation of reference toward any particular annotator's own referential preferences, annotators were told to err on the side of categorizing a reference as ambiguous if the pronoun could be interpreted as plausibly coreferential with either event participant, even if their own biases suggested a particular one. The number of cases in which pronouns were judged as ambiguous were spread relatively evenly across all four conditions; representative examples are shown in (9):

(9) a. The editor corrected the reporter who had made an obvious error. He was upset.

b. The nun praised the parishioner who was from San Diego county. She was excited.

c. The actress assisted the maid who needed help moving the furniture. She cleaned the dust bunnies under the couch.

The Cohen's kappa coefficient for measuring interannotator agreement was 0.71 for coherence and 0.88 for reference. In both annotations, the majority of disagreements were the result of one annotator rating a completion as ambiguous and the other not.

Of the 960 total continuations, we excluded those that did not mention the subject or object directly (e.g., mentioning a different entity, using a pleonastic subject, or mentioning both subject and object with plural *they*; 52 continuations, all in the FreePrompt condition), those in which reference was judged by at least one annotator to be ambiguous or for which the annotators picked

different referents (63 continuations), and those that were unanalyzable due to being ungrammatical or not complete sentences (4 continuations), leaving 841 continuations for analysis (398 FreePrompt and 443 PronounPrompt continuations). Of those, the referring expressions produced in the FreePrompt condition consisted of pronouns (*He, She*) and non-pronominal *the-NPs*. Of the 125 *the-NPs*, 112 used same NP as in context sentence (*the maid* → *the maid*), 6 used a shortened NP with the same head noun (*the baseball player* → *the player*), 5 used an NP with a different head noun (*the second grader* → *the student*), and 2 used an NP with a different head noun along with a modifier (*the rock star* → *the poor sap*).

Outcomes were modeled using mixed-effects logistic regression with maximal random effects structure when supported by the data [6]. Where a model did not converge, we successively removed random effects, chosen by the lowest variance. All factors were centered. To test for main effects and interactions, we conducted likelihood ratio tests between mixed-effects models differing only in the presence or absence of the targeted fixed effect.

3.3. Results

All of the predictions outlined at the top of this section were confirmed. Recall that the first prediction is that we would see a greater percentage of Explanation relations in the NoExplanationRC condition than in the ExplanationRC condition. This expectation is based on the idea that participants are less likely to write a completion that explains the event described by the matrix clause if an explanation was already provided by the RC. The results, shown in Figure 3, confirm the hypothesis (main effect of RC type on coherence relation, see Table 1).

This leads to the second hypothesis, which is that for the FreePrompt condition, we will see a greater percentage of next mentions of the object referent in the NoExplanationRC condition than in the ExplanationRC condition. This expectation results from the fact that the strong next-mention bias associated with object-biased IC verbs is conditioned on the next sentence providing an

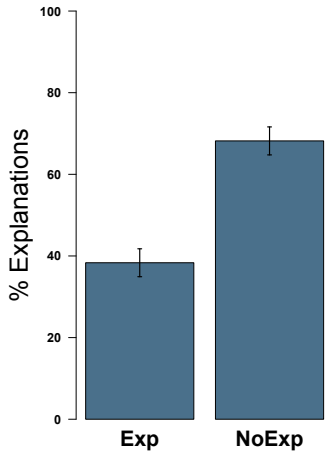

Figure 3: Percentage of Explanation Continuations in each RC condition (Standard Errors over Participant Means)

Fixed Effect	Coefficient	Std Error	z-value	p-value
RC	2.06	0.39	5.240	<0.001
Prompt	0.16	0.24	0.67	0.51
RCxPrompt	-0.03	0.66	-0.04	0.97

Table 1: Output for model of binary coherence outcome, Explanation or not, where convergence was reached with full random effect structure

explanation; thus fewer explanations should result in fewer object mentions. The results, shown in Figure 4, confirmed this as well (main effect of RC type on next mention for free-prompt subset, see Table 2).

Our third hypothesis is that the rate of pronominalization during production in the FreePrompt condition will *not* be similarly affected by RC condition; instead, it should only be affected by grammatical role, favoring pronominalization of the subject referent. This prediction results from the fact that we only expect a shift in coherence relations to affect the prior, to which production biases should be insensitive. Figure 5 shows the percentage of the time that

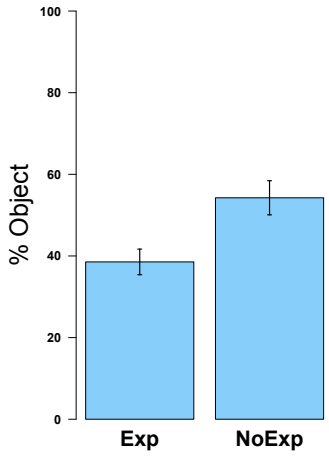

Figure 4: Percentage of Next-Mentions of Direct Object in each RC condition in FreePrompt Condition

references to a particular entity (subject or object) within a particular RC condition were pronominalized. The results confirm the effect of grammatical role and lack of interaction with RC condition (see Table 3).

Lastly, the analysis makes two predictions for pronoun interpretation biases $P(\textit{referent} \mid \textit{pronoun})$ as measured by the PronounPrompt condition. The results are shown in Figure 6. (Note that the bars for the FreePrompt condition, which represent the prior $P(\textit{referent})$, are identical to those shown in Figure 4.) First, we predict a greater percentage of object mentions in the NoExplanationRC condition than in the ExplanationRC condition, as a result of interpretation biases shifting along with the effect of RC type on the prior term in equation 3. This was confirmed (main effect of RC type on next mention for pronoun-prompt subset, see Table 2). Second, we predict a lower percentage of object mentions in the PronounPrompt condition than the FreePrompt condition, as a result of interpretation biases shifting along with the influence of grammatical role on the likelihood term in equation 3, specifically toward the subject. This was also borne out (main effect of prompt on next mention, see Table 2). The results therefore confirm that a subtle pragmatic manipulation

Fixed Effect	Coefficient	Std Error	z-value	p-value
Analysis of full data set				
RC	1.22	0.40	3.07	<0.005
Prompt	-1.27	0.34	-4.14	<0.001
RCxPrompt	0.85	0.43	1.64	0.08
Follow-up analysis of FreePrompt subset				
RC	0.72	0.42	2.43	<0.05
Follow-up analysis of PronounPrompt subset				
RC	1.17	0.43	3.09	<0.005

Table 2: Output for model of binary coreference outcome, object referent or not, where convergence was reached with full random effect structure

Fixed Effect	Coefficient	Std Error	z-value	p-value
RC	0.94	0.42	2.34	0.08
GramRole	4.11	0.60	6.94	<0.001
RCxGramRole	0.12	0.99	0.09	0.92

Table 3: Output for model of binary pronominalization outcome, pronoun or not, where convergence was reached with inclusion only of by-participant random slopes for grammatical role and the RC x grammatical role interaction and a by-item random slope for the interaction

of the context – whether or not an RC invites the inference of a cause of the matrix event – influences pronoun interpretation biases, but not production biases. The effect on interpretation is a surprise for theories based on surface-level characteristics of the context which, as noted earlier, find little to distinguish contexts like (8a-b).

3.4. Model Comparison

Finally, we can use the data collected here to test our Bayesian Model, i.e., that equation (3) captures the relationship between pronoun production and interpretation biases. We compare the predictions of the Bayesian Model

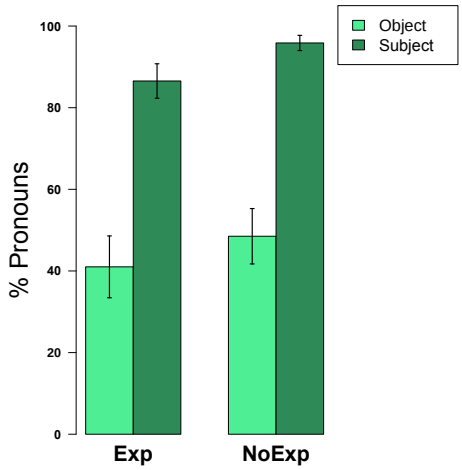

Figure 5: Rate of Pronominalization of Next-Mentioned Entity in FreePrompt Condition, by Grammatical Role and RC Condition

against two competing models. In each case, the predictions of the models can be evaluated using the results of our passage completion experiments. That is, the values on the right hand side of the formulas (specifically, the terms representing the next-mention and production biases) can be estimated using biases measured in the FreePrompt condition. The predicted interpretation biases of each model can then be compared against the actual interpretation biases measured in the PronounPrompt condition.

The first competing model is what we have called the Mirror Model, according to which the interpretation bias toward a referent is proportional to the likelihood that a speaker would produce a pronoun to refer to that referent. Capturing the intuition that hearers will base their interpretation decisions on their estimates about what referent mentions the speaker is likely to choose a pronoun to carry out, the predicted interpretation bias for this model is estimated using the pronominalization rate $P(\textit{pronoun} \mid \textit{referent})$ measured in the FreePrompt condition, normalized by the sum of the pronominalization rate for all referents. Interestingly, this formula essentially boils down to equation (3)

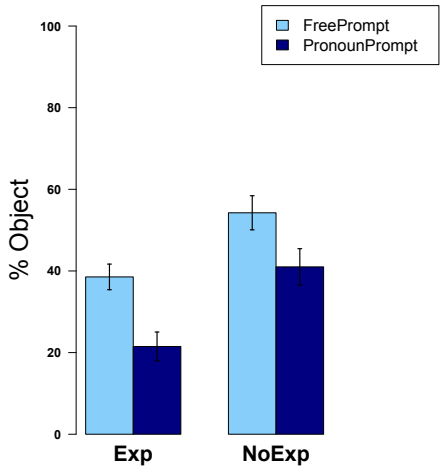

Figure 6: Percentage of Next-Mentions of the Direct Object by Prompt Condition and RC Condition

without the prior terms:[4]

$$(10)\quad P(\textit{referent} \mid \textit{pronoun}) \leftarrow \frac{P(\textit{pronoun} \mid \textit{referent})}{\sum_{\textit{referent} \in \textit{referents}} P(\textit{pronoun} \mid \textit{referent})}$$

The second competing model is what we have called the Expectancy Model, according to which the interpretation bias toward a referent equals the probability that the referent gets re-mentioned [3]. The predicted interpretation bias for this model is thus estimated to be the next-mention bias $P(\textit{referent})$ measured in the FreePrompt condition. Interestingly, this formula essentially boils down to equation (3) without the likelihood terms:[5]

$$(11)\quad P(\textit{referent} \mid \textit{pronoun}) \leftarrow \frac{P(\textit{referent})}{\sum_{\textit{referent} \in \textit{referents}} P(\textit{referent})}$$

[4] We use the assignment operator ($\leftarrow$) to underscore the fact that the model is not consistent with normative probability theory.

[5] Strictly speaking, the numerator of (11) is already a probability distribution, so there is no need for a normalization term. We write it like this to highlight the parallelism with equation (10) in effectively representing one half of the Bayesian formulation.

Finally, per equation (3), the predicted interpretation bias for the Bayesian Model results from combining the probabilities utilized by each of these other two models. We compare the predicted interpretation values for all three models against the observed pronoun interpretation biases $P(\textit{referent} \mid \textit{pronoun})$, as measured by the data collected in the PronounPrompt conditions.

	Actual	Bayesian	Mirror	Expectancy
ExplRC	.215	.229	.321	.385
NoExplRC	.410	.373	.334	.542
R^2 (participants/items)		.48/.49	.34/.42	.14/.12

Table 4: Actual and Predicted Rates of Pronominal Reference to Object

Table 4 shows the observed and model-predicted rates at which a pronoun prompt was interpreted to refer to the object across the two RC conditions. As can be seen, the observed values are most closely matched by the Bayes-derived values.

Finally, we also expect the Bayesian Model to exhibit a closer fit to the observed data with respect to the referential behavior both individual participants and individual items. Therefore, following Rohde and Kehler [37], we also test the degree of correlation between the values observed in the PronounPrompt condition and those generated by the three different analyses using linear models. The correlation is performed over participant and item means; each participant (or item) contributes a value for the four pronoun interpretation estimates in each of the RC condition referent combinations. We excluded from all models data points from participants (or items) for which the Mirror- and Bayes-derived values could not be estimated—specifically if a participants no-pronoun prompt responses for a particular RC condition contained no mentions of a particular referent or no pronouns for either referent; in both cases computing the predicted probabilities of the Mirror and Bayesian Models would involve division by zero. Each of the 40 participants stood to contribute observed- and model-derived probabilities for subject and object mention in each of the RC conditions; out

of this hypothetical total of 160 data points, 16 were excluded. Likewise for the 24 items, out of the hypothetical total of 96 datapoints, 6 were excluded. We expect that the predictions of all models will reveal some degree of correlation with the observed data: The Mirror Model should capture the differences in biases between subject and non-subject referents, whereas the Expectancy Model should capture differences across context type. Crucially, however, in combining the biases captured by both models, we expect the Bayesian Model to be more highly correlated than either of the other models alone. And this is in fact the case, as can be seen in the last row of Table 4.

4. Conclusion

A fundamental assumption that has underlied much of the literature on pronoun use is that there is a unified notion of prominence that mediates between production and interpretation. As we have noted, however, there has also been research that casts doubt on this assumption. We sought to examine the assumption by manipulating a more subtle aspect of the context than has been utilized in previous work, particularly whether or not an RC invites the inference of a cause of the event described by the matrix clause. Whereas many previous accounts offer no reason to think that such a manipulation would have an effect on pronoun interpretation biases, our experimental results show that it does, in exactly the respect predicted by the Bayesian Model. At the same time, we also found that the manipulation had no effect on pronoun production behavior. Whereas this mismatch may seem unintuitive, it reflects precisely the dissociation between production and interpretation captured by the Bayesian Model.

A key aspect of the design was the degree of indirection between the linguistic locus of the manipulation and the actual linguistic phenomenon being evaluated. As we indicated in Section 2.4, the analysis predicts that anything that affects a comprehender's prior expectations about what entities will be mentioned next will also affect pronoun interpretation biases. In the experiment

presented here, an RC manipulation was shown to affect interpretation biases by way of a specific hierarchical chain of causal dependencies: The invitation of an inference of an explanation directly affected the distribution of ensuing coherence relations, which in turn affected next-mention expectations, which in turn affected pronoun interpretation. So does this mean that future studies should put RC type alongside other factors when performing statistical analyses over pronoun interpretation data? The answer is clearly no. On our analysis, the pronoun model is not where this factor, nor others that are commonly included in such data analyses, properly belong. Instead, per Figure 2, this particular factor belongs in a model for predicting the distribution of ensuing coherence relations in a discourse context. The fact that RC type ultimately affects pronoun interpretation will be captured by the fact that models of interpretation will properly include next-mention expectations, and models of next-mention expectations will properly include coherence relation expectations. So whereas canonical 'bag-of-cues' analyses typically model the different factors that affect interpretation as an undifferentiated set, we would claim that the factors on the right-hand side of Figure 2 are properly modeled independently of pronouns. As far as pronoun interpretation is considered, the factors that influence the prior reside in a black box.

The results also show that whereas prior next-mention expectations are an important component of pronoun interpretation [3], they also confirm that interpretation biases are not reducible to those expectations. Our results thus bear relevance to studies on communicative efficiency such as that of Tily and Piantadosi [45], which appeal to the oft-noted Zipfian tendency for predictable information to be articulatorily reduced. Whereas there can be no doubt that there is a relationship between pronouns (in their role as the most reduced forms of referring) and referent predictability, our results again reveal that the connection is only partial. That is, pronouns do not merely function as unbound variables to be immediately assigned to the most expected entity at the time in the discourse at which they are encountered; they carry their own linguistic biases aside from entity predictability as captured by the likelihood term of

equation 3. Therefore, studies that examine pronominalization as a mechanism for languages to optimize communicative efficiency need to take this level of indirection into account.

A remaining question is how well the Bayesian Model captures the relationship between production and interpretation for a broader set of referring expression types across languages. Whereas we have focused on English personal pronouns here and in our prior work [25, 26], we would expect the weak version of the Bayesian Model to apply more generally across referential systems: If comprehenders reverse-engineer the referential intentions of the speaker by combining their estimates of production and next-mention biases when interpreting personal pronouns, the same should be true for other reference types as well. Importantly, however, we would expect that the relative impact of these two biases to vary across referential forms, since according to equation 3, the more strongly the production bias for a particular referring expression singles out a particular entity, the less easily next-mention biases will be able to push the preferred assignment away from that entity during interpretation.[6] Kaiser [23], for instance, reports that the interpretation of German demonstratives, which tend to be strongly biased toward the preceding object, is modulated to some degree by coherence relations, but less so than are personal pronouns, which tend to be biased toward the preceding subject. (For related work on demonstratives see [24] for Finnish and [39, 40] for German.) Such a finding is consistent with the Bayesian Model if in fact demonstratives have a stronger production bias toward the object than personal pronouns do toward the subject. A comprehensive evaluation of the Bayesian Model on demonstratives and other forms of reference awaits future work.

Finally, we return to the core question of the place of prominence in a theory

[6] To take an extreme example, consider first-person singular pronouns which, assuming a direct speech context, can only be used to refer to the speaker. In this case, the production bias for referents other than the speaker will be zero, and the speaker will be the predicted referent no matter how strongly the prior is biased toward other entities.

of pronoun interpretation, a relationship which, frankly, has historically been problematic. Whereas almost any paper on the topic will claim that pronouns refer to prominent (salient, accessible, predictable, in-focus) entities, it must be asked how the factors that contribute to prominence are determined. Unfortunately, the answer is typically that they are identified by researchers who examine the properties of pronominal referents in psycholinguistic and/or corpus data. This obviously leads to circularity: For the claim that pronouns refer to prominent entities to carry any content, the prominence of entities must be identifiable on grounds that are completely divorced from the forms of referring expression speakers use to refer to them.[7]

Only after examining production and interpretation behavior independently does it become clear that there can be no single notion of prominence that mediates between the two. According to the Bayesian Model, there are in fact two relevant notions. The first is entity predictability, as captured by the prior, and estimated by analyzing next-mention biases in free prompt passage completions. Crucially, the question of what entity will be mentioned next is independent of what particular referring expression a speaker chooses when referring to it, and hence as a notion of prominence it does not suffer from the circularity issues just noted. The second notion we need is the one that captures production biases. Is there a similarly reference-independent notion of prominence to be found here? Various authors [17, 18, 37, inter alia] have argued that the relevant linguistic notion is topicality, according to which the linguistic function of pronouns is to signal a continuation of the current topic. Insofar as this idea is correct and further that topicality can be defined as a coherent

[7]To be clear, we are not saying that previous work fails to make a distinction between a conceptually-independent notion of prominence or activation of entities in the discourse on the one hand, and the signals that referring expressions encode with respect to the notion of prominence or activation on the other. Indeed, many theories do just that [1, 21, 18, 28, inter alia]. The point is that the notion of prominence or activation appealed to is typically highly informed by the behavior of pronominal reference as witnessed in data, rather than being measurable by independent means.

linguistic notion on its own terms [19, 35] – both far from settled questions – there is hope that both types of prominence to which pronoun behavior is sensitive can be defined on independent grounds.[8]

The results presented here thus join the two studies utilizing IC contexts described in Rohde and Kehler [37] in showing that the predictions of the Bayesian Model provide a better fit with measured interpretation biases than competing models. (See also Rohde [36, pp. 126-128] for a discussion of the model with respect to transfer-of-possession contexts, which provide another context type in which the next-mention bias points strongly away from the non-subject.) Time will tell if the strong form of the Bayesian Model will hold up to further experimental scrutiny when applied to a broader range of context types and conditions. Whichever way that goes, we hope that our attempt to propose and defend a specific model will be part of a shift in emphasis from demonstrations that various linguistic factors have an effect on pronoun use to explanatory models that make their own quantitative, empirically testable predictions. Whereas empirical findings of the relevance of different factors on reference are useful,

[8]To be sure, there is evidence for the existence of topics provided by the behavior of linguistic phenomena other than pronouns. For instance, Strawson [44] famously argued that only topical definite NPs carry existential presuppositions, i.e., whereas the sentence *The King of France is bald* cannot be assigned a truth value due to the non-existence of the topical king, the sentence *Fred's class was visited by the King of France*, under a construal in which *Fred's class* is the topic, is more readily judged as simply false according to the intuitions of many speakers. Likewise, Kuno [30] proposes a Topichood Constraint on Extraction that captures why the question *Who did you read a book about?* is felicitous but *Who did you lose a book about?* is odd, as one could construe John Irving as being the topic of the sentence *I read a book about John Irving*, but not of *I lost a book about John Irving*, since one typically reads, but does not lose, books because of their content. Finally, Reinhart proposes the "Speaking of X" test for topichood, which captures why *Speaking of John Irving, I read a book about him* is felicitous whereas *Speaking of John Irving, I lost a book about him* is odd. See Gundel and Fretheim [20] for additional examples and useful discussion. Whereas these facts show that the evidence for topicality extends beyond the behavior of pronouns, there is no well-understood set of necessary and sufficient conditions on offer for identifying the topic across sentences and contexts.

they are not themselves theories; they are instead the data that we build theories to try to explain. If indeed the Bayesian Model fails to extend to a larger set of psycholinguistic examinations, we look forward to new quantitatively testable models offered as competitors, and to how those models inform questions about the relevant notions of prominence in reference and pragmatics more generally.

5. Acknowledgments

This paper is an expanded version of work reported on in Kehler and Rohde [27]. We thank research assistants Melodie Yen and Ksenia Kozhukhovskaya for their help in annotating data, as well as numerous university, workshop, and conference audiences that have provided helpful feedback.

[1] Ariel, M., 1990. Accessing Noun Phrase Antecedents. Routledge.

[2] Arnold, J.E., 1998. Reference Form and Discourse Patterns. Ph.D. thesis. Stanford University.

[3] Arnold, J.E., 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. Discourse Processes 21, 137–162.

[4] Arnold, J.E., 2010. How speakers refer: The role of accessibility. Language and Linguistic Compass 4, 187–203.

[5] Au, T.K., 1986. A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. Journal of Memory and Language 25, 104–122.

[6] Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language 68, 255–278.

[7] Bott, O., Solstad, T., 2012. The mechanics of causal interpretation: Explaining the implicit causality bias. Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing.

[8] Brown, R., Fish, D., 1983. The psychological causality implicit in language. Cognition 14, 237–273.

[9] Caramazza, A., Grober, E., Garvey, C., Yates, J., 1977. Comprehension of anaphoric pronouns. Journal of Verbal Learning and Verbal Behaviour 16, 601–609.

[10] Carreiras, M., Clifton, C., 1999. Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. Memory and Cognition 27, 826–833.

[11] Crawley, R.A., Stevenson, R.J., Kleinman, D., 1990. The use of heuristic strategies in the interpretation of pronouns. Journal of Psycholinguistic Research 19, 245–264.

[12] Fernandez, E., 2003. Bilingual Sentence Processing: Relative clause attachment in bilinguals and monolinguals. John Benjamins, Amsterdam.

[13] Frazier, L., Clifton, C., 1996. Construal. MIT Press, Cambridge, Mass.

[14] Fukumura, K., van Gompel, R.P.G., 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? Journal of Memory and Language 62, 52–66.

[15] Garvey, C., Caramazza, A., 1974. Implicit causality in verbs. Linguistic Inquiry 5, 459–464.

[16] Gibson, E., Piantadosi, S., Fedorenko, K., 2011. Using mechanichal turk to obtain and analyze english acceptability judgments. Language and Linguistics Compass 5, 509–524.

[17] Givón, T., 1983. Topic continuity in discourse: An introduction, in: Givón, T. (Ed.), Topic Continuity in Discourse: A Quantitative Cross-Language Study. John Benjamins, pp. 1–42.

[18] Grosz, B.J., Joshi, A.K., Weinstein, S., 1995. Centering: A framework for modelling the local coherence of discourse. Computational Linguistics 21, 203–225.

[19] Gundel, J., 1974. The Role of Topic and Comment in Linguistic Theory. Ph.D. thesis. University of Texas at Austin. Reprinted in Outstanding Dissertations in Linguistics Series, Garland Publishers, 1989.

[20] Gundel, J.K., Fretheim, T., 2004. Topic and focus, in: Horn, L.R., Ward, G. (Eds.), The Handbook of Pragmatics. Oxford: Basil Blackwell, pp. 175–196.

[21] Gundel, J.K., Hedberg, N., Zacharski, R., 1993. Cognitive status and the form of referring expressions in discourse. Language 69, 274–307.

[22] Hobbs, J.R., 1979. Coherence and coreference. Cognitive Science 3, 67–90.

[23] Kaiser, E., 2011. On the relation between coherence relations and anaphoric demonstratives in German, in: Proceedings of Sinn & Bedeutung 15, Saarland University Press. pp. 337–351.

[24] Kaiser, E., Trueswell, J., 2008. Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. Language and Cognitive Processes 23, 708–748.

[25] Kehler, A., Kertz, L., Rohde, H., Elman, J.L., 2008. Coherence and coreference revisited. Journal of Semantics 25, 1–44.

[26] Kehler, A., Rohde, H., 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. Theoretical Linguistics 39, 1–37.

[27] Kehler, A., Rohde, H., 2015. Pronominal reference and pragmatic enrichment: A bayesian account, in: Proceedings of the 37th Annual Conference of the Cognitive Science Society, Pasadena, CA. pp. 1063–1068.

[28] Kibrik, A.A., 2011. Reference in Discourse. Oxford University Press, Oxford.

[29] Koornneef, A.W., van Berkum, J.J.A., 2006. On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye-tracking. Journal of Memory and Language 54, 445–465.

[30] Kuno, S., 1987. Functional Syntax - Anaphora, Discourse and Empathy. The University of Chicago Press, Chicago and London.

[31] Levy, R., Jaeger, T.F., 2007. Speakers optimize information density through syntactic reduction, in: Proceedings of the 20th Conference on Neural Information Processing Systems (NIPS), p. 849856.

[32] McKoon, G., Greene, S.B., Ratcliff, R., 1993. Discourse models, pronoun resolution, and the implicit causality of verbs. Journal of Experimental Psychology 19, 1040–1052.

[33] Miltsakaki, E., 2007. A rethink of the relationship between salience and anaphora resolution, in: Branco, A. (Ed.), Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium, Lago, Portugal. pp. 91–96.

[34] Munro, R., Bethard, S., Kuperman, V., Lai, V., Melnick, R., Potts, C., Schnoebelen, T., Tily, H., 2010. Crowdsourcing and language studies: the new generation of linguistic data, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics. pp. 122–130.

[35] Reinhart, T., 1981. Pragmatics and linguistics: An analysis of sentence topics. Philosophica 27, 53–94.

[36] Rohde, H., 2008. Coherence-Driven Effects in Sentence and Discourse Processing. Ph.D. thesis. UC San Diego.

[37] Rohde, H., Kehler, A., 2014. Grammatical and information-structural influences on pronoun production. Language, Cognition and Neuroscience 29, 912–927.

[38] Rohde, H., Levy, R., Kehler, A., 2011. Anticipating explanations in relative clause processing. Cognition 118, 339–358.

[39] Schumacher, P.B., Dangl, M., Uzun, E., 2016. Thematic role as prominence cue during pronoun resolution in German, in: Empirical Perspectives on Anaphora Resolution. de Gruyter, Berlin, pp. 121–147.

[40] Schumacher, P.B., Roberts, L., Järvikivi, J., 2017. Agentivity drives real-time pronoun resolution: Evidence from German *er* and *der*. Lingua 185, 25–41.

[41] Simner, J., Pickering, M.J., 2005. Planning causes and consequences in discourse. Journal of Memory and Language 52, 226–239.

[42] Smyth, R.J., 1994. Grammatical determinants of ambiguous pronoun resolution. Journal of Psycholinguistic Research 23, 197–229.

[43] Stevenson, R., Crawley, R., Kleinman, D., 1994. Thematic roles, focusing and the representation of events. Language and Cognitive Processes 9, 519–548.

[44] Strawson, P.F., 1964. Identifying reference and truth-values. Theoria 30, 96–118.

[45] Tily, H., Piantadosi, S.T., 2009. Refer efficiently: Use less informative expressions for more predictable meanings, in: Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference.

Appendix: Stimuli

1. The actress assisted the maid who needed help moving the furniture / who worked the morning shift. (She)
2. The agent pities the rockstar who never seems to draw a crowd / who he managed to sign last week. (He)

3. The airline executive trusted the pilot who has a spotless flying record / who flies the nonstop from Philadelphia to San Diego. (He)
4. The announcer mocked the baseball player who had struck out four times in a row / who was first in the batting lineup. (He)
5. The babysitter despises the child who acts out at every opportunity / who is left home every Friday night. (She)
6. The businessman worships the accountant who found him a new set of tax deductions / who joined the team last week. (He)
7. The chessmaster ridiculed the novice who made a bad move / who had his lesson on Friday afternoons. (He)
8. The child fears the karate instructor who acts like a drill sargeant / who recently joined the training studio. (He)
9. The company president pacified the investor who was complaining about the impending merger / who stopped by for an impromptu meeting. (He)
10. The doctor reproached the patient who never takes her medicine / who came in at 3pm. (She)
11. The editor corrected the reporter who had made an obvious error / who had written that day's lead story. (She)
12. The employee envied the manager who had gotten a big raise / who worked in the next department over. (He)
13. The landlord valued the tenant who always keeps the apartment clean / who moved in last February. (He)
14. The musician detests the record label representative who always puts down her music / who flew into town this morning. (She)
15. The nun praised the parishioner who had worked so hard to help the poor / who was from San Diego county. (She)
16. The onlooker complimented the bride who looked dazzling in her designer wedding dress / who arrived at the restaurant. (She)
17. The politician resented the journalist who opened the interview by asking about her extramarital affairs / who interviewed her yesterday. (She)

18. The professor rewarded the student who received a perfect score on the final / who stopped by during office hours yesterday. (She)
19. The restaurant owner scolded the chef who was routinely letting food go to waste / who was hired last month. (He)
20. The scientist hates the health and safety expert who fines them after every inspection / who the company hired last month. (She)
21. The secretary adores the lawyer who gives her a big raise each year / who works in the office across from her desk. (She)
22. The teacher blamed the second grader who has a reputation for stealing things / who sits in the front of the class. (He)
23. The trainer consoled the boxer who had lost the fight / who had trained with him for 16 months. (He)
24. The wedding planner criticized the florist who had brought dying flowers to the ceremony / who had been hired for the ceremony. (She)